Statistical Methods in Medical Research

http://smm.sagepub.com/

Spatial health effects analysis with uncertain residential locations

Brian J Reich, Howard H Chang and Matthew J Strickland

Stat Methods Med Res published online 2 May 2012
DOI: 10.1177/0962280212447151

The online version of this article can be found at:
http://smm.sagepub.com/content/early/2012/05/02/0962280212447151

A more recent version of this article was published on - Mar 20, 2014

Published by:
SAGE
http://www.sagepublications.com

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–13
© The Author(s) 2012
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280212447151
smm.sagepub.com
SAGE

Spatial health effects analysis with uncertain residential locations

Brian J Reich,[1] Howard H Chang[2] and Matthew J Strickland[3]

Abstract

Spatial epidemiology has benefited greatly from advances in geographic information system technology, which permits extensive study of associations between various health responses and a wide array of socio-economic and environmental factors. However, many spatial epidemiological datasets have missing values for a substantial proportion of spatial variables, such as the census tract of residence of study participants. The standard approach is to discard these observations and analyze only complete observations. In this article, we propose a new hierarchical Bayesian spatial model to handle missing observation locations. Our model utilizes all available information to learn about the missing locations and propagates uncertainty about the missing locations throughout the model. We show via a simulation study that this method can lead to more efficient epidemiological analysis. The method is applied to a study of the relationship between fine particulate matter and birth outcomes is southeast Georgia, where we find smaller posterior variance for most parameters using our missing data model compared to the standard complete case model.

Keywords

Bayesian hierarchical model, conditionally autoregressive prior, data imputation, geographic information system, missing data

1 Introduction

Spatial epidemiology has benefited greatly from advances in geographic information system (GIS) technology. GIS facilitates the study of associations between various health responses and a wide array of socio-economic and environmental factors. Despite improvements in data collection methods and GIS technology,[1,2] missing data remain prevalent in spatial health datasets. In this article, we specifically address the problem of study participants with uncertain residential locations.

[1]Department of Statistics, North Carolina State University, USA
[2]Department of Biostatistics and Bioinformatics, Emory University, USA
[3]Department of Environmental Health, Emory University, USA

Corresponding author:
Brian J Reich, Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
Email: brian_reich@ncsu.edu

This study is motivated by an analysis of the effects of fine particulate matter on preterm birth and low birth weight in southern Georgia. Figure 1 shows the 24 counties in the study domain as well as the 177 census tracts in these mostly rural counties. While the majority of air pollution and health studies have been conducted in urban communities, there is a growing interest to examine the effects of air pollution in rural communities. This is because rural communities often have a different susceptible population composition and are exposed to a different pollution mixture. In our dataset, several predictors are known at the census tract level, including the variable of interest, fine particulate matter exposure, and the tract's median income. The county of residence is known for all mothers; however, for 9.7% of the mothers in the study, the census tract of residence is unknown (ranging by county from 2.7% to 57.1%). This uncertainty in the spatial location poses challenges in the health model, as it is not clear how to assign exposure to these mothers or how to model variability in risk across tracts. The ability to utilize health data with uncertain residential location is particularly important in this setting because missing geocodes are most prevalent in rural regions, which have considerably smaller sample size compared to urban communities.

There is an extensive literature on missing data methodology, for example, Little and Rubin[3] and Enders.[4] However, relatively few missing data methods are available that are specifically tailored for spatial data. Recently, methods for preferential sampling have been developed for cases where the locations of the observations are selected in a way that depends on spatial distribution of the response.[5,6] In the models for these data, the locations are modeled jointly with the response. Although the spatial locations are modeled statistically, unlike our application the datasets considered in preferential sampling do not have missing values. In a similar work, Reich and Bandyopadhyay[7] model spatial data with predefined measurement locations, but with missing responses. In this approach, the location of missing responses are allowed to depend on the true value of the spatial process being measured. The situation addressed here differs from Reich and Bandyopadhyay[7] in that the location of the observations is missing, but not the response.

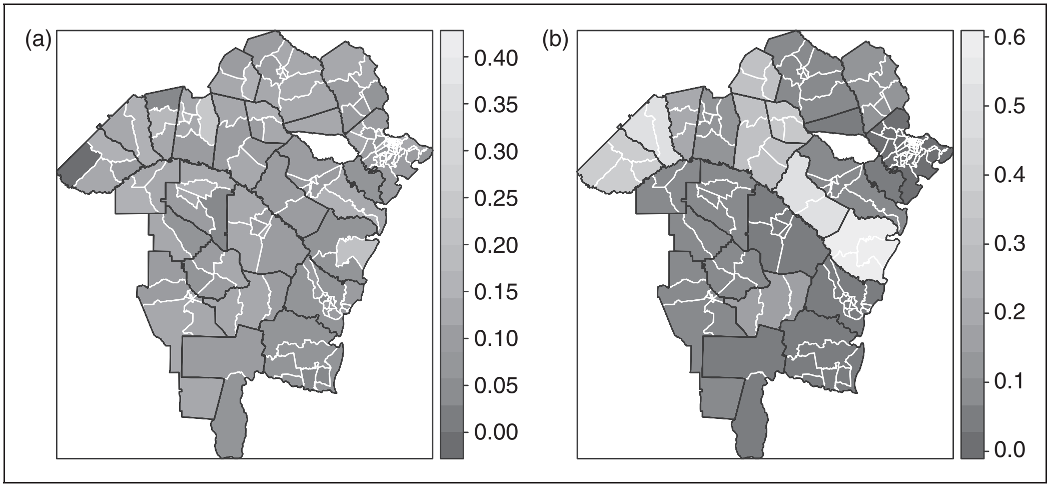

Figure 1. Plot of the counties and census tracts in southern Georgia that constitute the study domain. (a) preterm birth rate and (b) proportion of missing tracts by county.
The white areas indicate tracts with no births during the study period.

Most similar to our method is that of Cressie and Kornak,[8] who consider the case of a continuous spatial domain and discuss the effects of adjusting for measurement error in the sampling locations in the usual geostatical/kriging setup. Our case is different in that we are dealing with areal data, and thus, the missing locations are restricted to a finite set of indices, and we use covariates to inform about the missing data process.

In this article, we propose a hierarchical Bayesian approach that accounts for uncertainty in the residential location of the study participants. We treat the missing locations as unknown random variables in the Bayesian model. By modeling these data hierarchically, we exploit all available information to inform about the missing locations, including the mother's characteristics (e.g. race and martial status), census information (e.g. the percent married and racial composition of the census tracts), and the proportion of missing data in each census tract. By modeling the data using Bayesian methods, uncertainty about the missing locations is naturally propagated throughout the model, including the posteriors of the parameter relating fine particulate matter with the health response, which is the primary interest. The proposed approach also has the advantage of being a relatively straightforward addition to standard computational algorithms used for Bayesian spatial epidemiology models. In particular, it can be implemented using the popular software `OpenBUGS`.

The remainder of the article proceeds as follows. Section 2 describes the motivating data. The statistical model and corresponding computational details are given in Sections 3 through 5. In Section 6, we conduct a simulation study to illustrate the benefits of the proposed approach compared to the standard method. The method is applied to Georgia birth data in Section 7. Section 8 concludes.

2 Description of the Georgia birth data

Birth record data were obtained from the Office of Health Indicators for Planning, Georgia Division of Public Health. The study region consists of 24 counties (177 census tracts) located in southeastern Georgia (Figure 1). We considered only singleton births without structural birth defects conceived from the period 1 January 2002 to 31 December 2005. Gestational age was defined as the number of completed weeks between the reported date of last menstrual period and the date of birth. We removed records with birth weight less than 400 g and records with gestational age less than 26 weeks or greater than 44 weeks. We restricted the analysis to non-Hispanic white and non-Hispanic black mothers between the age of 15 and 44. The data are summarized in Table 1.

Daily ambient concentrations of fine particulate matter ($PM_{2.5}$) were obtained from the Statistically Fused Air Quality[9] database (FSD) available at the EPA website, http://www.epa.gov/heasd/sources/projects/CDC/index.html. The FSD database contains predicted daily $PM_{2.5}$ concentrations over contiguous $12 \times 12\,\text{km}^2$ grid cells. The predictions are based on a Bayesian space–time hierarchical model that combines (1) observed $PM_{2.5}$ levels from the monitoring network and (2) deterministic outputs from the Models-3/Community Multiscale Air Quality model. To account for spatial misalignment between the grid cells and census tract boundaries, we first calculated the proportion of the tract area that fall within each FSD grid cell. Then, daily tract-level values were obtained by taking a weighted average.

The following tract-level population statistics were obtained from Census 2000 for females between the age of 15 and 44: total population count, proportion of black race, and the mean and SD of the age distribution. The median personal income in 1999 and the proportion who were married were also obtained for females, without the age restriction, due to limited census data.

Table 1. Summary statistics of the Georgia birth data.

	Missing tract-level geocode	
	No ($N = 40{,}963$)	Yes ($N = 4391$)
Preterm birth (%)	11.3	11.2
Low birth weight (%)	6.9	7.1
Female infant (%)	49.0	50.3
Married (%)	57.4	55.0
Tobacco use (%)	11.8	14.4
Maternal age		
Mean	25.5	24.8
SD	5.7	5.5
Mother's race (%)		
Black	35.0	27.7
White	65.0	72.3
Mother's education (%)		
Some college or higher	43.9	31.7
High school or lower	56.1	68.3
Conception season (%)		
March–May	25.1	25.5
June–August	23.9	23.6
September–November	25.4	25.2
December–February	25.6	25.7

3 Hierarchical model for missing spatial locations

Let Y_i be the binary indicator of low birth weight or preterm birth for birth $i = 1, \ldots, n$ and $s_i \in \{1, \ldots, N\}$ the index of the residential region (e.g. census tract) for observation i. We use the binary regression model

$$g[\mathrm{P}(Y_i = 1)] = \alpha(s_i) + \mathbf{X}_i^T \boldsymbol{\gamma} + \mathbf{W}(s_i)^T \boldsymbol{\omega} + \mathbf{Z}_i(s_i)^T \boldsymbol{\beta} \tag{1}$$

where g is a link function (e.g. logit or probit), $\boldsymbol{\alpha} = [\alpha(1), \ldots, \alpha(N)]^T$ the spatial random effects to control for unmeasured spatial baseline risks, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_w)^T$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ the fixed effects. We separate the predictors into three types: $\mathbf{X}_i$ are descriptions of the ith birth and are known even in the absence of s_i (e.g. the mother's age or education attainment), $\mathbf{W}(s_i)$ descriptions of the mother's residential region (e.g. the tract's median income) and are the same for all mothers in region s_i, and $\mathbf{Z}_i(s_i)$ covariates that depend on the mother's region and vary for mothers in region s_i (e.g. the average $\mathrm{PM}_{2.5}$ concentration in region s_i over mother i's first trimester).

Our focus is to develop a statistical model for the case where a significant proportion of the residential locations are missing. Let δ_i be the binary indicator of a missing value. For the majority of observations, GIS calculates s_i accurately and thus $\delta_i = 0$. For some observations, GIS is unable to specify s_i and thus $\delta_i = 1$. Fortunately, in most such cases, we are able to place s_i into a subset of indices $S_i \subset \{1, \ldots, N\}$, e.g. S_i may be the indices of the census tracts within a county if the mother's residential county can be obtained directly from the birth certificate or by geocoding using ZIP code. Clearly, this complicates the spatial model, as it is unclear how to assign these observations a spatial random effect. This uncertainty also affects the regression portion of the model, since some of the predictors depend on s_i.

We account for uncertainty in the residential region using Bayesian modeling. The spatial index s_i is treated as an unknown parameter in the hierarchical model. To exploit all available information that can be used to impute the missing locations, we specify the joint distribution of all data for observation i, $[Y_i, \delta_i, \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i]$. To specify the model, without loss of generality, we define the joint distribution as the product of three conditional distributions

$$[Y_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i, s_i] = [Y_i, \delta_i | \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i][\mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i | s_i][s_i] \tag{2}$$

As is often the case in missing data modeling, assumptions are required about the missing data mechanism. A key assumption in our analysis is that given the spatial location and the predictors, the birth outcome is independent of the missing data indictor, i.e.

$$[Y_i, \delta_i | \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i] = [Y_i | \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i][\delta_i | \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i] \tag{3}$$

This seems reasonable in our setting because many GIS errors are caused by unmatched addresses in geocoding, for example, due to missing street numbers, and many variables known to be associated with coding errors[10] are provided by the birth records and included in $\mathbf{X}_i$.

The models for the three conditional distributions in equation (2) are described below. The model for $Y_i | \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i$ is given by equation (1). A flexible model for the missing data mechanism $[\delta_i | \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i]$ is

$$g[\mathrm{P}(\delta_i = 1 | \mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i, s_i)] = a(s_i) + \mathbf{X}_i^T\mathbf{b} + \mathbf{W}(s_i)^T\mathbf{c} \tag{4}$$

where $\mathbf{a} = [a(1), \ldots, a(N)]^T$ are spatial random effects and $\mathbf{b}$ and $\mathbf{c}$ regression parameters. We exclude time-varying coefficients $\mathbf{Z}_i$ from this model, since it seems unreasonable that GIS success rate depends on time and is correlated with covariates such as ambient air pollution.

In some cases, there may be prior information about the GIS success rate in each region which could be incorporated in the prior for $\mathbf{a}$. However, considerable effort is required to assess GIS error rate, especially for a large study region.[11] In the absence of prior information, it will be difficult to estimate the spatial random effects $\mathbf{a}$. Therefore, to avoid identifiability issues, we omit these random effects from the model, giving

$$g[\mathrm{P}(\delta_i = 1 | \mathbf{X}_i, \mathbf{W}, s_i)] = \mathbf{X}_i^T\mathbf{b} + \mathbf{W}(s_i)^T\mathbf{c} \tag{5}$$

This probability depends on s_i through $\mathbf{W}(s_i)$ and therefore helps impute the missing indices. Also, intuitively, the data provide information about $\mathbf{c}$ by comparing the proportion of missing tracts in each county with the average $\mathbf{W}$ over census tracts in the county. For these reasons, we select equation (5) as the missing data model in Section 7.

Next, we describe the model for $[\mathbf{X}_i, \mathbf{W}, \mathbf{Z}_i | s_i]$. The spatial covariate $\mathbf{W}$ is considered fixed and thus not modeled stochastically. For Georgia data, there are no missing values or uncertainty in $\mathbf{X}_i$. However, it is still important to model their distribution given s_i because this provides information to impute missing census tracts. For example, a married mother is more likely to reside in a census tract with high proportion of married women. Therefore, a model for the proportion of married women in each census tract is needed so that it can be combined with maternal information to help impute the missing census tract indicators. The predictors $\mathbf{X}_i$ are a mix of continuous and binary variables. If the jth covariate, X_{ji}, is continuous, we model $X_{ji}|s_i \sim \mathrm{N}[\mu_j(s_i), \sigma_j^2(s_i)]$; if X_{ji} is binary, we model $P(X_{ji} = 1|s_i) = \mu_j(s_i)$. In our data, there is a single exposure (i.e. $q = 1$), which is modeled as $Z_i|s_i \sim \mathrm{N}[\mu_{0i}(s_i), \sigma_{0i}^2(s_i)]$. If the exposure is known exactly in each census tract, then $\sigma_{0i}^2(s_i) = 0$ and $\mu_{0i}(s_i)$ is the exposure for tract s_i. In most air pollution studies, there is uncertainty in the exposure in each tract due to measurement error and incomplete sampling. Our approach in dealing with the uncertainty via priors for μ_{ji} and σ_{ji} is described in Section 4. It is also straightforward to treat $[\mathbf{X}_i|s_i]$ as a multinomial distribution where each outcome represents a unique strata of $\mathbf{X}_i$. This is particularly useful when the prior is derived from census tables of several variables.

The third component of equation (2) is the prior of the residential location. We assume that $100\pi_j\%$ of the mother's reside in region j, with $\sum_{j=1}^{N} \pi_j = 1$. For partially observed locations restricted to $s_i \in S_i$, the conditional prior is $\mathrm{Prob}(s_i = j\,|s_i \in S_i) = \pi_j/(\sum_{k \in S_i} \pi_k)$. The probabilities $(\pi_1, \ldots, \pi_N)$ may be fixed based on census data or given prior to account for uncertainty in population density.

4 Prior distributions

Spatial random effects $\boldsymbol{\alpha}$ in equation (1) are modeled using a conditionally autoregressive (CAR) model.[12] The CAR model can be defined by its full conditional distributions

$$\alpha_j|\alpha_l \text{ for all } \; l \neq j \; \sim \mathrm{N}\left[\mu + \frac{\rho}{m_j}\sum_{l \sim j}(\alpha_l - \mu), \frac{\tau^2}{m_j}\right] \tag{6}$$

where $l \sim j$ indicates that regions l and j are adjacent and m_j is the number of regions that are adjacent to region j. The CAR prior has three parameters: μ is the location, τ^2 controls the variance, and $\rho \in [0, 1]$ controls the degree of spatial dependence. Combining these full conditionals gives a multivariate normal joint distribution for $\boldsymbol{\alpha}$ with mean $(\mu, \ldots, \mu)^T$ and covariance $\tau^2(\mathbf{M} - \rho\,\mathbf{D})^{-1}$, where $\mathbf{M}$ is the diagonal matrix with diagonal elements $\{m_1, \ldots, m_N\}$, $\mathbf{D}$ the adjacency matrix with (j, l) element equal to $I(j\sim l)$, and I the binary indicator function. We denote this model $\boldsymbol{\alpha} \sim \mathrm{CAR}(\mu, \tau, \rho)$. For the data analysis in Section 7, we fix $\rho = 1$, which gives an improper intrinsic prior. For simulated data in Section 6, we fixed $\rho = 0.9$, since it is not possible to simulate data from the improper model with $\rho = 1$. To complete the health model, we specify uninformative priors $\mu, \gamma_j, \omega_j, \beta_j \overset{iid}{\sim} \mathrm{N}(0,\ 10^2)$ and $\tau^{-2} \sim \mathrm{gamma}(0.1,\ 0.1)$. Similarly, the prior for the missing data parameters are b_j, $c_j \sim \mathrm{N}(0,\ 10^2)$.

The prior mean and variance of the air pollution exposure, μ_{0i} and σ_{0i}, are provided by the FSD database described in Section 2. Although there is substantial variation in the daily exposures, the uncertainty is negligible after averaging over the entire first trimester and is therefore ignored by setting $\sigma_0 = 0$. The remaining values of μ_j, σ_j, and $\pi_1, \ldots, \pi_N$ are fixed based on census data.

5 Computational details

Although it would be straightforward to implement in standard software such as `OpenBUGS`,[13] we perform Markov chain Monte Carlo (MCMC) sampling using `R`.[14] We use a probit link $g(x) = \Phi(x)$, where Φ is the standard normal distribution function, for both the health model (1) and missing data model (5). Therefore, we introduce two auxiliary variables[15] for each subject, U_i and V_i, with

$$\begin{aligned} U_i &\sim \mathrm{N}[\alpha(s_i) + \mathbf{X}_i^T\boldsymbol{\gamma} + \mathbf{W}(s_i)^T\boldsymbol{\omega} + \mathbf{Z}_i(s_i)^T\boldsymbol{\beta}, 1] \\ V_i &\sim \mathrm{N}[\mathbf{X}_i^T\mathbf{b} + \mathbf{W}(s_i)^T\mathbf{c}, 1] \end{aligned} \tag{7}$$

The auxiliary variables relate to the responses by $Y_i = I(U_i > 0)$ and $\delta_i = I(V_i > 0)$. Marginalizing over the auxiliary variables gives the desired probabilities in equations (1) and (5). After introducing these latent variables, all model parameters are conjugate, and therefore, Gibbs sampling is used to generate samples from the posterior distribution.

Most of the full conditionals needed for Gibbs sampling follow from standard conjugacy relationships. A few parameters have non-standard full conditionals, which are specified below. The auxiliary variables have truncated normal full conditionals

$$\begin{aligned} U_i|\text{rest} &\sim \mathrm{TN}_{\mathcal{D}(Y_i)}[\alpha(s_i) + \mathbf{X}_i^T\boldsymbol{\gamma} + \mathbf{W}(s_i)^T\boldsymbol{\omega} + \mathbf{Z}_i(s_i)^T\boldsymbol{\beta}, 1] \\ V_i|\text{rest} &\sim \mathrm{TN}_{\mathcal{D}(\delta_i)}[\mathbf{X}_i^T\mathbf{b} + \mathbf{W}(s_i)^T\mathbf{c}, 1]. \end{aligned} \tag{8}$$

where the truncation region is $\mathcal{D}(Y) = [0, \infty]$ if $Y = 1$ and $\mathcal{D}(Y) = [-\infty, 0]$ if $Y = 0$. Conditioned on the latent variables in equation (7) and assuming Gaussian priors, the full conditionals of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\omega}$, $\boldsymbol{\beta}$, $\mathbf{b}$, and $\mathbf{c}$ follow from the usual normal/normal conjugacy relationship and are thus omitted. The CAR variance τ^2 has the standard gamma full conditional from normal/gamma conjugacy and is omitted.

The final full conditional needed for the Gibbs sampler is for the missing location indices s_i for observations with $\delta_i = 1$. The full conditional is

$$\begin{aligned} P(s_i = l\,|\text{rest}) &= \frac{\eta_{il}\pi_j \mathrm{I}(l \in S_i)}{\sum_{k=1}^N \eta_{ik}\pi_k \mathrm{I}(k \in S_i)} \quad \text{where} \\ \eta_{il} &= \phi[U_i|\alpha(l) + \mathbf{X}_i^T\boldsymbol{\gamma} + \mathbf{W}(l)^T\boldsymbol{\omega} + \mathbf{Z}_i(l)^T\boldsymbol{\beta}, 1] \;\times \phi[V_i|\mathbf{X}_i^T\mathbf{b} + \mathbf{W}(l)^T\mathbf{c}, 1] \\ &\quad \times \prod_{j=1}^{p_1} \phi[X_{ij}|\mu_j(l), \sigma_j^2(l)] \times \prod_{j=p_1+1}^{p} \mu_j(l)^{X_{ij}}[1 - \mu_j(l)]^{1-X_{ij}} \end{aligned} \tag{9}$$

The covariates are ordered so that the first p_1 covariates in $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ are Gaussian and the remaining $p - p_1$ are binary, and $\phi(y|m, s)$ is the $\mathrm{N}(m, s^2)$ density function.

For simulated data in Section 6, we generate 10,000 samples and discard the first 1000 as burn-in. For the data analysis in Section 7, we generate 25,000 samples and discard the first 10,000 as burn-in. Convergence is monitored using trace plots of several representative parameters. We find that convergence is almost immediate for this model.

6 Simulation study

The data are simulated on a 6×6 regular grid of $N = 36$ census tracts. The tracts are partitioned into a 3×3 grid of counties, as shown in Figure 2. Mothers' census tracts are randomly assigned with equal probability. The county is known for each mother, but the census tract is missing with probability π. In addition to air pollution exposure, there is a single ($p = 1$) binary predictor X_i, which is known for all mothers. We vary the mean of the predictor by alternating columns in order to allow contrast in prevalence within a particular county. The predictor is generated as

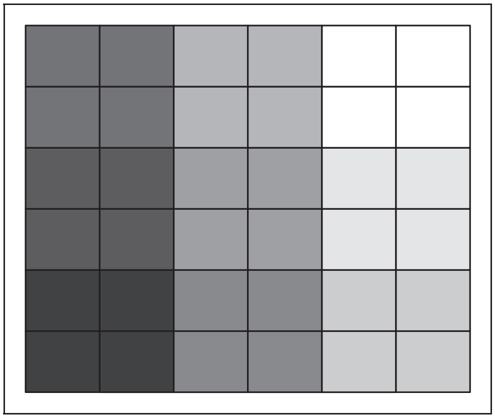

Figure 2. Plot of the census tracts for the simulation study. Shading corresponds to the tract's county.

$X_i|s_i \sim \text{Bernoulli}(\mu_X)$ for s_i in even numbered columns (Figure 2) and $X_i|s_i \sim \text{Bernoulli}(1-\mu_X)$ for s_i in odd numbered columns; we assume μ_X is known to replicate known census information. For each mother, we randomly draw an exposure for each census tract, $\mathbf{Z}_i = [Z_i(1), \ldots, Z_i(N)]^T \overset{iid}{\sim} \text{CAR}(0,1,\rho_z)$. For each simulated dataset, the spatial random effects are drawn as $\boldsymbol{\alpha} = [\alpha(1), \ldots, \alpha(N)]^T \sim \text{CAR}(\mu, \sigma^2, \rho)$. The responses are then generated as

$$\text{P}(Y_i = 1) = \Phi[\alpha(s_i) + X_i\omega + Z_i(s_i)\beta] \tag{10}$$

Each simulated dataset has $n = 1000$, $\omega = 0.5$, $\mu = -1.0$, $\sigma = 0.1$, and $\rho = 0.9$. For simplicity, we do not include spatial covariate $\mathbf{W}(s)$. For each simulated dataset, we fit three models:

(1) *Complete case (CC)*: discard all observations with a missing census tract.
(2) *Spatial uncertainty (SU)*: treat the missing census tracts as unknown parameters as described above.
(3) *Oracle (O)*: use all observations as if we knew the missing census tracts.

Model 1 is the simplest model which ignores uncertainty in the spatial locations by removing all problematic observations. Model 2 is our proposed approach. The final model cannot be fit to real data because it uses information that is not known to the analyst. This is included as a reference to gage the effectiveness of the data imputation method.

We take as the base case for the simulation $\pi = 0.2$, $\mu_X = 0.3$, $\beta = 0.5$, and $\rho_z = 0.9$. We then simulate data for several designs by varying these four factors. For each design we generate 200 datasets. Figure 3 plots root mean squared error (RMSE) for β, averaged over the simulated dataset. We also computed the empirical coverage probabilities. They were at or above the nominal level for all models and simulation designs, and therefore, we do not present them here.

As expected, the RMSE of the SU model is between the RMSE of the CC and O models for all settings. The improvement of the SU model compared to the usual CC model is highly dependent on the percent missing. The relative MSE of the SU to the CC model varies from $(0.053/0.056)^2 = 0.941$ with $\pi = 0.1$ to $(0.053/0.069)^2 = 0.700$ with $\pi = 0.4$. Increased variability in the mean of the covariate across census tracts (i.e. small μ_X) improves the ability of the model to impute the

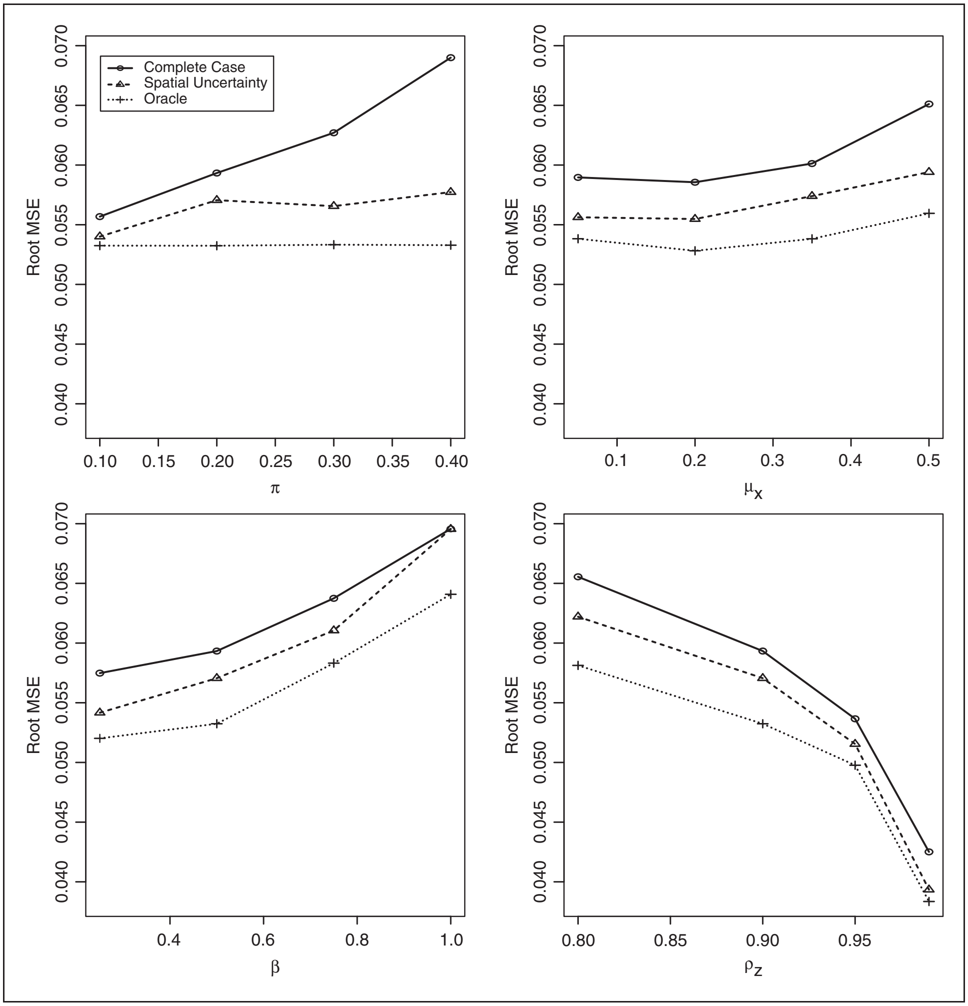

Figure 3. Simulation study results.
RMSE is plotted by the proportion of missing data (π), the mean of the covariate in the even numbered columns (μ_X), the exposure effect β, and the spatial dependence parameter of the air pollution exposure (ρ_z). The Monte Carlo standard error is between 0.0028 and 0.0056 for all RMSE estimates.
RMSE: root mean square error.

missing census tracts. The average (over missing observation and dataset) posterior probability on the correct tract increases from 0.26 with $\mu_X = 0.5$ to 0.43 with $\mu_X = 0.05$. Surprisingly, this does not translate into improved RMSE for β, as the relative MSE is fairly constant for all μ_X.

The SU model improves MSE for moderate signal with $\beta = 0.25$. However, for a very strong signal, the CC model is nearly as effective as the SU model. Finally, all methods give smaller MSE when the spatial dependence parameter for the air pollution exposure, ρ_Z, increases. The SU model becomes increasingly efficient as the dependence increases. The relative MSE of the SU to the CC model varies from $(0.058/0.066)^2 = 0.900$ with $\rho_Z = 0.80$ to $(0.038/0.043)^2 = 0.857$ with $\rho_Z = 0.99$. This is because uncertainty in the census tract is less problematic when the census tracts have similar exposures.

7 Analysis of Georgia birth data

We examined the associations between exposure to ambient $PM_{2.5}$ during the first trimester and the risk of low birth weight (less than 2500 g) and preterm birth (less than 37 gestational week) separately via a spatial probit regression model. The total study population is 45,354 births and the raw prevalence was 11.3% for preterm birth and 6.9% for low birth weight. First trimester exposure was defined as the average daily $PM_{2.5}$ concentrations during the first 13 weeks of pregnancy. The model included the following confounders from the birth records: maternal age, indicator for race, indicator for infant sex, indicator for marital status (married or unmarried), and indicators for some college or higher and self-reported tobacco use during pregnancy. To control for unmeasured temporal confounders, the model also included: (a) indicators for the season of conception (spring: March–May, summer: June–August, autumn: September–November, winter: December–February); and (b) a smooth function of conception date modeled using natural cubic splines with four degrees of freedom. To control for spatial confounders, we included tract-level median personal income from Census 2000 and tract-specific spatial random intercepts via the CAR specification.

Table 2 summarizes the posteriors of the coefficients in the health and missing-tract models. Results are given for the SU model and CC models defined in Section 6. Unmarried, black mothers without college education and with tobacco use have a higher risk of both preterm birth and low birth weight. Also, older mothers are at higher risk of preterm birth, and female babies are at higher risk of low

Table 2. Posterior mean (SD) for the Georgia birth outcomes analysis.

	Preterm birth		Low birth weight	
	CC	SU	CC	SU
(a) Health model				
$PM_{2.5}$	−0.0005 (0.0053)	0.0022 (0.0056)	−0.0043 (0.0060)	−0.0034 (0.0057)
Age	0.005 (0.002)	0.005 (0.002)	0.002 (0.002)	0.002 (0.002)
Married	−0.075 (0.021)	−0.074 (0.019)	−0.102 (0.024)	−0.105 (0.022)
Black	0.224 (0.020)	0.218 (0.019)	0.385 (0.023)	0.385 (0.022)
Female infant	−0.023 (0.016)	−0.028 (0.016)	0.159 (0.020)	0.161 (0.019)
Tobacco use	0.181 (0.025)	0.189 (0.024)	0.367 (0.028)	0.380 (0.026)
Education	−0.084 (0.019)	−0.080 (0.018)	−0.093 (0.023)	−0.090 (0.022)
Spring	0.057 (0.024)	0.054 (0.022)	0.044 (0.028)	0.035 (0.026)
Summer	0.036 (0.024)	0.036 (0.023)	0.019 (0.028)	0.005 (0.027)
Fall	0.040 (0.024)	0.032 (0.023)	0.017 (0.028)	0.009 (0.026)
Median income	0.027 (0.032)	0.033 (0.031)	−0.035 (0.035)	−0.028 (0.032)
(b) Missing tract model				
Mother's age	0.002 (0.002)		0.002 (0.002)	
Married	−0.017 (0.020)		−0.016 (0.020)	
Black	−0.318 (0.021)		−0.317 (0.021)	
Female infant	0.020 (0.017)		0.021 (0.017)	
Tobacco use	−0.009 (0.025)		−0.009 (0.025)	
Education	−0.183 (0.019)		−0.183 (0.019)	
Median income	−0.831 (0.034)		−0.803 (0.034)	

CC: complete case; SU: spatial uncertainity.
Results are given for separate analysis of preterm birth and low birth weight, and for CC and SU models.

birth weight. Based on these data, there is no statistically significant relationship between $PM_{2.5}$ and either low birth weight or preterm birth.

The effect estimates are similar for both the SU and CC models. As expected, the posterior SD is smaller for the most parameters under the SU model compared to the CC model. For the confounding predictors (age–education in Table 2) for preterm birth, the ratio of posterior variance for the CC compared to SU model ranges from $(0.0199/0.0192)^2 = 1.07$ for maternal race to $(0.0205/0.0191)^2 = 1.15$ for maternal marital status. For low birth weight, the ratio of posterior variances ranges from 1.00 for maternal age to $(0.0230/0.0215)^2 = 1.14$ for maternal education. These predictors are known for all observations, even those with missing census tracts, and so, the SU model provides smaller variance due to the larger sample size. Curiously, the posterior SD of the

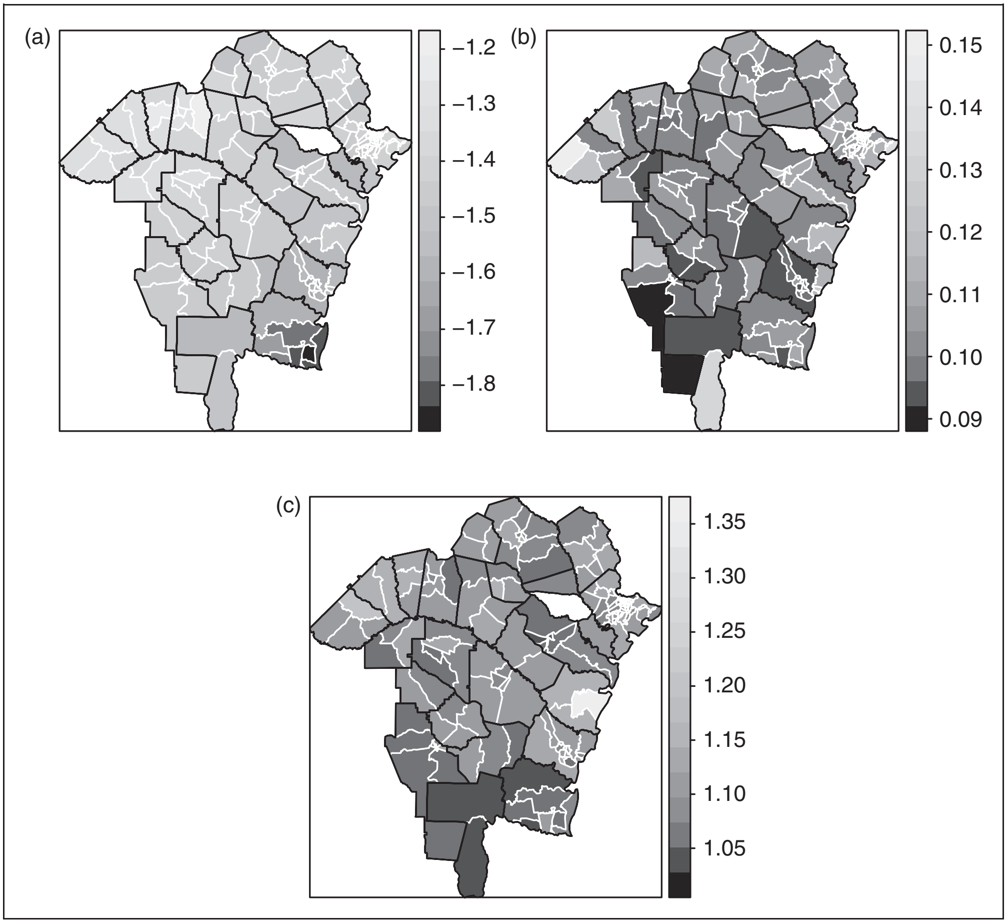

Figure 4. Summary of the spatial random effects α for the preterm birth analysis. (a) posterior mean; (b) posterion SD; and (c) relative variance of CC versus SU.
Panels (a) and (b) plot the posterior mean and SD of α for the SU model and panel (c) plots the ratio of posterior variance for the CC model relative to the SU model. CC: complete case; SU: spatial uncertainity.

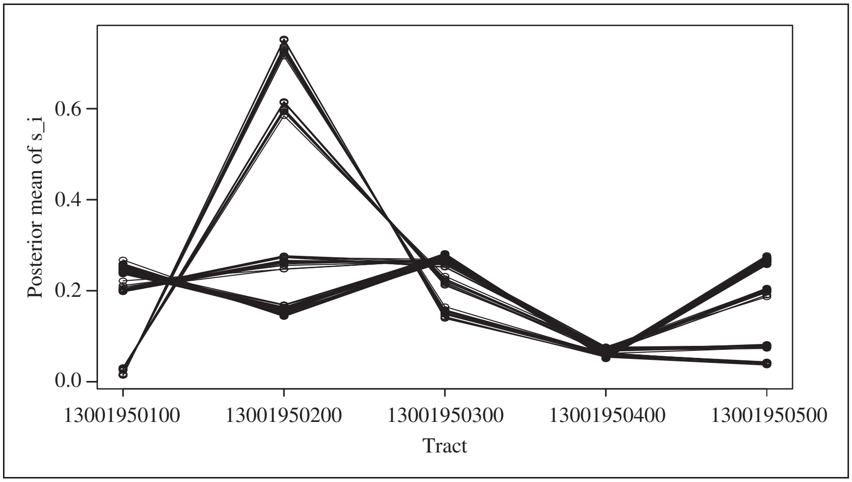

Figure 5. Posterior distribution from the preterm birth analysis for the census tract index of each mother with a missing census tract from Appling County.
Each set of five connected points represents the posterior distribution of s_i for one mother.

$PM_{2.5}$ effect for preterm birth is larger for the SU than the CC model. This may be because the increased sample size for the SU is offset by uncertainty in the $PM_{2.5}$ predictor.

The spatial random effect estimates in Figure 4 show considerable spatial variation, with significantly lower risk in the southeast portion of the spatial domain. Figure 4(c) plots the ratio of posterior variance from the CC to the SU model. As with the fixed effects, the variance of the random effects is considerably smaller for the SU model.

There are several statistically significant predictors of the missing census tracts (Table 2). The results are nearly identical for preterm birth of low birth weight because the majority of information for these parameters comes from the regression of the missing data indicators and covariates, and these are the same for both health responses. White mothers without college education and in tracts with a low median income are more likely to have missing tract information. This provides information about the missing census tract, as does the mother's covariate information. Figure 5 shows the posterior distribution of the missing tract index for each mother from Appling County. For all mothers, the probability for tract 13001950400 is small, since this tract has the smallest population (9% of the county's population). The posterior probability for tract 13001950200 varies dramatically across mothers. The main driver of the probability is the mother's race and marital status; of the five tracts in Appling County, this tract has the highest proportion of black residents (50.5%, no other tract exceeds 25%) and lowest proportion of married mothers (59.3%, all other tracts exceed 66%).

8 Discussion

In this article, we develop a hierarchical Bayesian model which incorporates uncertainty about the spatial location of the study participants. The method is straightforward to implement in standard MCMC algorithms. We show via simulation that properly accounting for SU can lead to a substantial improvement in parameter estimation over the standard approach of discarding incomplete observations. The method is then applied to a study of the association between fine particulate matter and birth outcomes in Georgia. Although this analysis did not reveal a statistically significant association, accounting for SU reduced the posterior variance of the coefficients by as much as 15%. Our simulation study suggests that reductions in variance will be even larger for studies with more missing spatial information. For this large study where the CC model already gives small standard errors, the credible sets that do or do not include zero are the same for both the CC and SU models. However, in more smaller studies, a 15% reduction in posterior variance could reveal new environmental factors relating to birth outcomes and lead to public health initiatives to improve birth outcomes.

As with many approaches to missing data problems, our analysis relies on strong assumptions. In particular, we are assuming that after accounting for covariate effects, accurate recording of the spatial location is independent of the response. This seems reasonable in our setting, but this assumption should be carefully scrutinized in future applications. Technically, the more complicated case of missing spatial locations depending on the response is still considered missing at random, and therefore, it may be possible to include the response in the logistic regression model for the missing data indicator. This is an area of future work.

Funding

This study was partially supported by National Institutes of Health grant 2R01ES014843-04A1 and USEPA grant R834799. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

1. Henry K and Boscoe F. Estimating the accuracy of geographical imputation. *Int J Health Geogr* 2008; **7**: 1–10.
2. Goldberg D, Wilson J, Knobloch C, et al. An effective and efficient approach for manually improving geocoded data. *Int J Health Geogr* 2008; **7**: 1–10.
3. Little RJ and Rubin DB. *Statistical analysis with missing data*. Hoboken, NJ: Wiley, 2002.
4. Enders CK. *Applied missing data analysis (methodology in the social sciences)*. New York: The Guilford Press, 2010.
5. Diggle P, Meneze R and Su T. Geostatistical inference under preferential sampling (with discussion). *J Roy Stat Soc Ser C (Appl Stat)* 2010; **59**: 191–232.
6. Pati D, Reich BJ and Dunson DB. Bayesian geostatistical modeling with informative sampling locations. *Biometrika* 2011; **98**: 35–48.
7. Reich BJ and Bandyopadhyay D. A latent factor model for spatial data with informative missingness. *Ann Appl Stat* 2010; **4**: 439–459.
8. Cressie N and Kornak J. Spatial statistics in the presence of location error with and application to remote sensing of the environment. *Stat Sci* 2003; **18**: 436–456.
9. McMillan M, Holland D, Morara M, et al. Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* 2009; **21**: 48–65.
10. Kravets N and Hadden W. The accuracy of address coding and the effects of coding errors. *Health Place* 2007; **13**: 293–398.
11. Strickland M, Siffel C, Gardner B, et al. Quantifying geocode location error using gis methods. *Environ Health* 2007; **6**: 1–8.
12. Gelfand AE, Diggle PJ, Fuentes M, et al. *Handbook of spatial statistics*. Boca Raton, FL: Chapman & Hall/CRC, 2010.
13. Lunn D, Spiegelhalter D, Thomas A, et al. The BUGS project: evolution, critique, and future directions. *Stat Med* 2009; **28**: 3049–3067.
14. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2010. http://www.R-project.org/, ISBN 3-900051-07-0.
15. Albert JH and Chib S. Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* 1993; **88**: 669–679.